DOCUMENT RESUME

ED 453 248 TM 032 785

AUTHOR Roberts, J. Kyle; Onwuegbuzie, Anthony J.; Eby, J. Robert
TITLE The Introduction of a Measure of Instrument Homogeneity for Interpreting Low Reliability Coefficients.
PUB DATE 2001-04-14
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Error of Measurement; *Reliability; *Scores
IDENTIFIERS *Alpha Coefficient; *Homogeneity Tests; Monitoring

ABSTRACT

This paper suggests that although data from a homogenous sample might yield less reliable scores than did an inducted sample, these data should not be discarded until further examination of the data is conducted. The paper presents two statistics for monitoring data homogeneity and one statistic for correcting alpha when homogeneity is large. The first of the statistics for monitoring sample homogeneity is computed as the difference in mean item variances, expressed as a percentage. The second monitoring statistic is termed the "relative squared standard error of estimate index" (J. Roberts and A. Ongwuegbuzie, 2000). The new statistic for correcting alpha is developed for the situation in which the homogeneity is large and access to a previously normed sample is not available. The use of all three techniques is illustrated. (Contains 3 tables and 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: ALTERNATIVE APPROACHES FOR INTERPRETING ALPHA

ED 453 248

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. K. Roberts

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Introduction of a Measure of Instrument Homogeneity for Interpreting Low Reliability Coefficients

J. Kyle Roberts

Baylor College of Medicine

Anthony J. Onwuegbuzie

Valdosta State University

J. Robert Eby

Texas A&M University

TM032785

Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA, April 14, 2001.

Correspondence should be addressed to: J. Kyle Roberts, Center for Educational Outreach, Baylor College of Medicine, 1709 Dryden, Suite 545, Houston, TX 77030. E-Mail: jroberts@bcm.tmc.edu, 713-798-7673

BEST COPY AVAILABLE

Abstract

Much of the current research concerning reliability is emphatically suggesting that researchers gather their own reliability estimates when administering an instrument. It has also been recommended that data with low reliability then be discarded. While some data obtained from instruments that originally yielded reliable results may be unreliable, it does not necessarily follow that the data are unuseful to researchers. This paper will contend that although data from a homogeneous sample might yield less reliable scores than did an inducted sample, these data should not be discarded until further examination of the data is conducted. The authors will present two statistics for monitoring data homogeneity and one statistic for correcting alpha when homogeneity is large.

The Introduction of a Measure of Instrument Homogeneity for Interpreting Low Reliability Coefficients

Much of the current literature concerning reliability is emphatically suggesting that researchers obtain their own reliability estimates when gathering new data from a previously-developed instrument and then report these estimates (Vacha-Haase, Kogan, & Thompson, in press) and confidence intervals (Onwuegbuzie & Daniel, 2000) in their final reports. Moreover, Thompson and Vacha-Haase (2000), encouraged researchers not to "induct" reliability estimates for their own dataset from previous studies, but to obtain reliability estimates for their own dataset because reliability estimates are affected by individual sample characteristics. Concerning this practice of "inducting" reliability estimates, Pedhazur and Schmelkin (1991) stated:

> [Reliability estimates printed in test manuals] may be useful for comparative purposes, but it is imperative to recognize that the relevant reliability estimate is the one obtained for the sample used in the [current] study under consideration. (p. 86)

Similarly, Dawis (1987) contended that although the type of instrument utilized may influence reliability, reliability also might be influenced by sample composition and sample variability. Therefore, it is imperative that researchers compute and interpret reliability coefficients for their underlying sample.

Noting the problems with interpreting alpha as a consistent measure of reliability, Pedhazur and Schmelkin (1991) stated, "coefficient alpha will underestimate the reliability of a measure when its items are not at least essentially tau-equivalent" (p. 100). Pedhazur and

Schmelkin (1991) then give a brief discussion of the problems associated with interpreting alpha with differing test lengths (number of items), with restricting time allotted to administer the test, and when instrument homogeneity (unidimensionality) is high. The last of these problems, homogeneity, is the focus of the present essay.

Coefficient Alpha

For the purposes of this paper, we will use the formula for coefficient alpha that was originally developed by Cronbach (1951). For reference purposes, the formula for computing alpha is

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_x^2}\right) \qquad (1)$$

where k is the number of items, $\Sigma\sigma_i^2$ is the sum of the individual test item variances, and σ_x^2 is the total test variance. Although typically Kuder Richardson-20 is used with dichotomously-scored items, it should be noted that alpha equals KR-20 because each kth $p_k q_k$ equals each ith σ_i^2, and across all k items $\Sigma p_k q_k = \Sigma\sigma_i^2$ (Thompson, 1999). Consulting Equation 1, a large alpha would be yielded from scores that necessarily had a small sum of individual item variances and a large total test variance. Likewise, a dataset yielding a small alpha coefficient would be produced by scores that had high individual item variances and a small total test variance. It should also be noted that although conceptually alpha represents a squared metric, it is mathematically possible to obtain an alpha value less than zero. This occurs when the sum of the individual item variances exceeds the total test variance.

Consider the following heuristic example of datasets that would yield a low reliability estimate. The first example is a measure given in which all of the students simply guessed at responses to items and the variance of the individual items was as large or even larger than the

total test variance. This might be the case if an instrument like the Scholastic Assessment Test was given to second graders. The second example is when there is a lack of variability among examinees. For example, if a second-grade spelling test was administered to college English majors, all of the examinees would probably score close to the same value, thus yielding a small alpha.

It has been recommended consistently in the literature that scores yielding low reliability estimates either be considered extremely suspect or discarded (Abelson, 1997). While in many instances, scores that yield low reliability coefficients indicate poor psychometric properties, it does not necessarily follow that the underlying data are always "unuseful" to researchers. Consider an example when a depression measure is administered to students who are all relatively not depressed, that is, who represent a homogeneous sample with respect to depression scores. Such scores likely would yield a low reliability estimate. However, interpreting this reliability coefficient without taking into consideration the homogeneous nature of the sample would be misleading. Rather, what is needed in such instances is not to discard the data, but to examine the scores further to determine why low reliability estimates were obtained from an instrument that produced high reliability estimates in the original test sample. In an attempt to overcome problems associated with low reliability estimates in homogeneous samples, Magnusson (1967) offered a formula based on the sample variance from the original normative group (i.e., the inducted sample) and the underlying sample group. Magnusson's (1967) formula can be used to predict the reliability of scores from a present sample, based on the reliability of scores from the inducted sample and the standard deviations of both samples.

Heuristic Example

For the purposes of the present article, two heuristic datasets were utilized that were both (hypothetically) derived from the same eight-item instrument. The first dataset was designed such that the scores reflected what a homogeneous sample might look like if drawn from a population of scores. It should be noted that there are slight variations within each item because including data in which all of the examinees had the same score contributes nothing to the variability of the data. The second dataset was generated to represent a completely random set of scores. The data used are illustrated in Tables 1 and 2. (We would like to bring to the reader's attention the fact that we are not advocating that such small samples be utilized, due to their ensuing low statistical power for detecting relationships. These small datasets are used for illustrative purposes only.)

Insert Tables 1 and 2 about here

Once data were generated, reliability analyses were conducted using the Statistical Package for the Social Sciences (SPSS; SPSS Inc., 2000). Results from the reliability analyses are provided in Table 3. The second column in Table 3 illustrates the variance and reliability estimates for the random data in Table 2. The third column represents the variance and reliability estimates for the homogeneous data in Table 1. In the fourth column, a hypothetical dataset also has been included to simulate plausible test manual data. We will use this dataset for comparative purposes with our other two random and homogeneous datasets.

Insert Table 3 about here

From Table 3, we see that the alpha coefficients from the random, homogeneous, and inducted dataset are .097, .071, and .883, respectively. In this example, although the inducted scores yielded a high reliability coefficient, scores from the two heuristic datasets yielded low reliability estimates. The reason that alpha on these two datasets is low is because the sum of the item variances is almost as great as the total test variance. In the inducted dataset, the sum of the item variances is small, whereas the total test variance is large, thereby yielding a large alpha coefficient.

The mean item variances in Table 3 are the overall mean of the eight item variances for each dataset. These values are .2408, .0588, and .1658 for the random, homogeneous, and inducted datasets, respectively. We see that the mean item variances effectively discriminate the random dataset from the homogeneous dataset. Specifically, whereas the mean item variance for the random dataset is larger than that for the inducted sample, the mean item variance for the homogeneous dataset is smaller than that for the inducted sample.

The variance of the mean item variances (Table 3) is simply the squared standard deviation (σ^2) of the mean item variances. These values are .0001, .0000, and .0037 for the random, homogeneous, and inducted datasets, respectively. It should be noted that a small variance of the mean item variances indicates that the mean item variances are relatively the same (e.g., either all small or all large). In each of the three datasets presented here, the variance of the mean item variances is small, thus indicating that nearly all of the item variances, within each dataset, are clustered relatively close together. Therefore, the variance of the mean item variances does not adequately discriminate the random and homogeneous samples.

Also presented in Table 3 is the ratio of the sum of the individual test item variances to the total test variance for the random, homogeneous, and inducted datasets. These values are .9132, .9414, and .2211, respectively. This ratio, which represents the last entry of Equation 1 above, is the most important component of an alpha coefficient. Interestingly, these ratios do not adequately discriminate the random dataset from the homogeneous dataset.

Also presented in Table 3 is the square of the standard error of measurement. The standard error of measurement, or the standard deviation of errors of measurement, provides an absolute rather than a relative measure of the extent to which raw and true scores are equivalent (Crocker & Algina, 1986). Although the true score can never be known, the standard error of measurement can be applied to an individual's observed score to set "plausible limits" for locating the true score. These "plausible limits" provide confidence bands for interpreting an obtained score. The smaller the standard error of measurement, the smaller the confidence band, and the greater the confidence that the observed score is near the true score. From Table 3, we see that the squared standard error of measurement effectively discriminates the random dataset from the homogeneous dataset. Specifically, whereas the squared standard error of measurement for the random dataset is larger than that for the inducted sample, the squared standard error of measurement for the homogeneous dataset is smaller than that for the inducted sample.

Statistics for Detecting Sample Homogeneity

Although Magnusson (1967) had previously defined methods for predicting alpha given psychometrics from original test manuals, not until recently have methods been developed for detecting sample homogeneity. Roberts and Onwuegbuzie (2000) have developed two statistics that investigate data homogeneity, given previous test psychometrics. The first of these statistics is computed as the difference in mean item variances. This statistic can be expressed as a

percentage to yield what they termed the "relative mean item variance index." The relative mean item variance index is computed as

$$\frac{M\sigma_u^2 - M\sigma_{u'}^2}{M\sigma_u^2} \tag{2}$$

where $M\sigma_u^2$ is the mean item variance for the inducted sample and $M\sigma_{u'}^2$ is the mean item variance for the underlying sample. This index ranges from $-\infty$ to 1, with positive values indicating that the study sample is more homogeneous than the inducted sample, and negative values indicating that the study sample is less homogeneous than is the original sample. This index can be expressed as a percentage by multiplying the index by 100. For the present sample, the random group has a relative mean item variance index of -.4524, whereas the index pertaining to the homogeneous dataset is .6454. The negative relative index associated with the random dataset suggests that the low reliability index is not the result of homogeneity. Conversely, the relative index value pertaining to the homogeneous dataset indicates that the low reliability coefficient obtained for the homogeneous dataset is explained to some extent by the relative homogeneous nature of the dataset. In this latter case, it could be argued that the low reliability coefficient represents a statistical artifact.

The second statistic developed by Roberts and Onwuegbuzie (2000) was termed the "relative squared standard error of estimate index," and can be computed as

$$\frac{s_u^2 - s_{u'}^2}{s_u^2} \tag{3}$$

where s_u^2 is the squared standard error of estimate from inducted sample and $s_{u'}^2$ is the squared standard error of estimate from underlying sample. This index also ranges from $-\infty$ to 1, with positive values indicating that the study sample is more homogeneous than is the inducted

sample, and negative values indicating that the study sample is less homogeneous than is the inducted sample. For the present heuristic example, the random group has a relative squared standard error of estimate index of -1.7140, whereas the index pertaining to the homogeneous dataset is .3382. As for the relative mean item variance index, the negative relative squared standard error of estimate index associated with the random dataset suggests that the low reliability index is not the result of homogeneity. Further discussion of these statistics is presented in Roberts and Onwuegbuzie (2000).

Alpha ROE for Dichotomously Scored Instruments

Although the relative mean item variance index and the relative squared standard error of estimate index are both useful statistics when investigating data obtained from a previously published instrument, it does not always stand to reason that test publishers will report item variances and/or standard error of the estimate for the inducted sample. Until now, there has been no way of correcting alpha in samples where homogeneity is large and access to a previously-normed sample is not available.

We define a new measure, which we will call α_{ROE}. With N = the number of test questions, n_i being the number of respondents to each test question, and a_i, b_i being the number of students who answered question i, correctly and incorrectly (respectively), we have

$$\alpha_{ROE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{a_i - b_i}{n_i}\right)^2. \qquad (4)$$

Since $a_i + b_i = n_i$ for each question, the we can substitute to get

$$\alpha_{ROE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2a_i - n_i}{n_i}\right)^2. \qquad (5)$$

In some sense, this is a variance measure, since

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{2a_i - n_i}{n_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2\left(a_i - \frac{n_i}{2}\right)}{n_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}2^2\left(\frac{a_i - \frac{n_i}{2}}{\frac{n_i}{2}}\right)^2 = \frac{4}{N}\sum_{i=1}^{N}\left(\frac{a_i - \frac{n_i}{2}}{\frac{n_i}{2}}\right)^2. \quad (6)$$

However, it is only natural to investigate what the range of this new α would be. Supposing that everyone answers every question with a response of "1", then $a_i = n_i \forall i$. Hence,

$$\alpha_{ROE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2a_i - n_i}{n_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2n_i - n_i}{n_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}1 = 1. \quad (7)$$

On the other hand, suppose that $a_i = b_i = \frac{n_i}{2}$ (i.e. each question has an equal number of people answering *a* and *b*). Then,

$$\alpha_{ROE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2a_i - n_i}{n_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2\frac{n_i}{2} - n_i}{n_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}0 = 0. \quad (8)$$

So we see that for a dataset with perfect homogeneity, α_{ROE} receives a value of "1", whereas with a perfectly heterogeneous dataset, α_{ROE} receives a value of "0". Thus we have the bounds of α_{ROE}. Therefore, in any given dataset, α_{ROE} is measuring the amount of variance from complete data heterogeneity, since $\frac{n_i}{2} = a$ in a dataset of complete heterogeneity. Thus, intuitively, we are measuring how far the obtained data is from complete dataset heterogeneity.

Applying this formula to the current "homogeneous" data set we can see that for each of the eight items, 1 person out of 16 has a score differing from the others (note: this was setup to provide some variance in the dataset so that Cronbach's alpha could be computed). Thus, for α_{ROE} for this dataset, we see that the square of the number of items correct for each item is 0.766 across all eight items. Therefore α_{ROE} equals 0.766.

Likewise, we can compute the α_{ROE} for the "random" dataset by applying the above formula. In doing so we have

$$\frac{\left(\frac{9-7}{16}\right)^2+\left(\frac{9-7}{16}\right)^2+\left(\frac{8-8}{16}\right)^2+\left(\frac{5-11}{16}\right)^2+\left(\frac{9-7}{16}\right)^2+\left(\frac{9-7}{16}\right)+\left(\frac{7-9}{16}\right)^2+\left(\frac{10-6}{16}\right)^2}{8} \quad (9)$$

or

$$\frac{.016+.016+0+.141+.016+.016+.016+.063}{8} \quad (10)$$

such that α_{ROE} is .036 for the "random" dataset.

When investigating datasets with low reliability, computing α_{ROE} seems to be the best and easiest means of investigating attenuation in reliability estimates due to data homogeneity. While the relative mean item variance index and the relative squared standard error of the estimate index helps determine the relative homogeneity based on the original inducted sample, α_{ROE} seems to be a more appropriate statistic for use when the test manual psychometrics are not available (or complete) and for use when no other reference dataset is available.

Because α_{ROE} is a squared metric, it has a lower bound of zero. And since the total number correct for any given item cannot exceed the number of respondents, the upper bound for this statistic is one. Consequentially, when consulting α_{ROE}, it can be interpreted in the same manner as Cronbach's alpha. For example, in the case where coefficient alpha were .80, we could state that at least 80% of the total test score variance is due to true score variance. In the case of α_{ROE}, if α_{ROE}= .80, we could say that at least 80% of the lack of total test score variance (and individual item variance) is due to examinee homogeneity.

Real-Life Examples Illustrating When Homogeneity Yields Meaningful and Worthless Low Reliability Coefficients

One of the main concerns posited in this paper is that researchers, when confronted with data that yield a low alpha coefficient, should seek to investigate the reason(s) for this low reliability estimate. We believe that in the case where the data are homogeneous, it does not follow that simply discarding the data is merited. Consider the two following examples of data that would yield low reliability estimates. The first example we will refer to as an illustration of "bad" homogeneity, and the second example we will refer to as an illustration of "good" homogeneity.

After administering an examination, a researcher discovers that the dataset she is investigating had low reliability because an achievement test had been administered to a group of sixth graders that was originally intended for second graders. As a result, the entire cohort of students achieved high scores on the examination. In this instance, it could be justifiably contended that this instrument yields unreliable data for this specific sample. As has been mentioned previously, this does not mean that the instrument is unreliable (instruments are neither reliable nor unreliable), but simply that the instrument is inappropriate for use with this cohort. This is an example of what is meant by "bad" homogeneity, or homogeneity that should consequentially lead to a discarding of the data.

The second example represents a different but equally realistic scenario. Suppose that a depression scale has been administered to 45 patients of an outpatient clinic for depression. Researchers have administered this instrument because they are concerned with the effectiveness of a certain intervention and are measuring the gain scores on the depression scale. In this case, it would be expected that the reliability estimates would be very small for the 45 patients who are

relatively homogeneous because they are in a depression clinic being treated for depression. However, simply disposing with the data in this instance seems unwarranted if indeed low reliability is due to homogeneity and not to randomness. The goal of administering the instrument to these patients is to monitor the effectiveness of the intervention. In research studies like this, the effect size will be maximized if and only if participants from one extreme of the scale (e.g., clinically-depressed participants) are the focus of the study. Otherwise, ceiling effects could confound results, thereby providing rival explanations (i.e., low internal validity). Additionally, findings from such an investigation would only be generalizable (i.e., have maximal external validity) if a clinically-depressed sample is utilized. This is an example of "good" homogeneity, or homogeneity that attenuates the reliability estimate in a manner that is directly a function of the level of homogeneity of the sample.

In the second example, it is extremely likely that the mean item variances were smaller than the mean item variances reported in the test manual. It is also very likely that α_{ROE} is very large since all examinees scored roughly the same on most of the administered items. If this were the case, the data should still be used in the analysis, particularly if the sample size was large, because the low reliability estimate is due to individual homogeneity and thus appears acceptable considering the context of the study. Moreover, with respect to the homogeneous dataset presented in Table 1, it should be noted that if only the reliability coefficient in Table 3 had been reported, and further examination of the reliability properties had not been performed, readers of the subsequent final report might not have looked favorably on scores that yielded a reliability coefficient of .071. However, adding an explanation of the reasons to the possibility of a small coefficient alpha, such as an α_{ROE} of .766, would strengthen the researchers argument for using the existing data for further analyses.

Conclusion

Although the major focus of this paper has been to advocate that researchers spend time examining the reasons behind data yielding low reliability estimates, it should be noted that, previously, no correction statistics existed for alpha coefficients with homogeneous samples. This can and should be the focus of future research in this area. While we encourage the investigation of examinee homogeneity through α_{ROE}, the mean item variances, and the squared standard error of measurement, no specific guidelines have been developed for determining what is and what is not an acceptable threshold for data homogeneity. However, as illustrated above, both α_{ROE} and the (squared) standard error of measurement provide extremely useful information about the degree of homogeneity of an underlying sample.

Because low reliability estimates reduce the statistical power associated with hypothesis tests (Onwuegbuzie & Daniel, 2000), researchers should utilize larger samples, when they expect that their sample is homogeneous, in order to compensate for the corresponding reliability-based loss in statistical power. In fact, increasing the sample size (i.e., a research-based consideration) would probably provide a better correction for attenuated relationships than any statistical correction of the reliability coefficient itself (i.e., a statistical consideration), just as randomizing participants to treatment conditions in an experimental design is superior to using analysis of covariance techniques or other statistical adjustments to analyze data stemming from non-experimental research designs.

Nevertheless, encouraging the investigation of coefficients of reliability can only help the growth of accountability in test usage. It is our contention that researchers should be encouraged not only to compute reliability estimates for their own data, but also to investigate and to explain

why the reliability estimates differ from the original test manual norming (i.e., inducted) sample. Such information would allow readers to put subsequent findings in a more appropriate context.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow & S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Chicago, IL: Holt, Rinehart, and Winston, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481-489.

Magnusson, D. (1967). *Test theory*. Boston, MA: Addison-Wesley.

Onwuegbuzie, A.J., & Daniel, L.G. (2000, November). Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. In G. Halpin (Chair), *Reliability Issues*. Symposium conducted at the annual conference of the Mid-South Educational Research Association, Bowling Green, KY.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Roberts, J. K., & Onwuegbuzie, A. J. (2000). *Alternative approaches for interpreting alpha with homogeneous subsamples.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY, November 16, 2000.

SPSS Inc. (12000). *SPSS 10.0 for Windows.* [Computer software]. Chicago, IL: SPSS Inc.

Thompson, B. (1999, February). *Understanding coefficient alpha, really.* Paper presented at the annual meeting of the Educational Research Exchange, College Station, TX.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174-195.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (in press). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*(4).

Table 1

Homogeneous dataset

Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1	1	0	0	1	1	0	0
1	1	0	0	1	1	0	0
1	1	0	0	1	1	0	0
1	0	0	0	1	1	0	0
1	1	0	0	1	1	0	0
1	1	0	0	1	1	0	0
0	1	0	0	1	1	0	0
1	1	0	0	1	1	0	0
1	1	0	1	1	1	0	0
1	1	0	0	0	1	0	0
1	1	1	0	1	1	0	0
1	1	0	0	1	0	0	0
1	1	0	0	1	1	1	0
1	1	0	0	1	1	0	0
1	1	0	0	1	1	0	1
1	1	0	0	1	1	0	0

Table 2

Random dataset

Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1	0	1	0	0	1	1	1
1	0	0	0	1	1	1	1
0	1	0	0	1	1	0	0
1	0	0	1	1	1	1	1
0	1	1	0	1	0	1	1
0	1	0	0	1	0	1	0
1	0	1	1	0	0	1	1
0	0	0	0	0	0	1	0
1	1	1	0	0	1	0	1
0	1	0	0	1	1	0	0
0	1	1	1	1	1	0	1
0	0	0	1	1	1	0	1
1	0	1	0	0	0	0	1
1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	0	1	1	0	1

Table 3

Item and Test Characteristics for the Two Heuristic Datasets and the "Inducted" Dataset

Variables	Random Dataset	Homogeneous Dataset	"Inducted" Dataset
Item 1 variance	0.2461	0.0588	0.1094
Item 2 variance	0.2461	0.0588	0.1875
Item 3 variance	0.2500	0.0588	0.2125
Item 4 variance	0.2113	0.0588	0.2148
Item 5 variance	0.2461	0.0588	0.2461
Item 6 variance	0.2461	0.0588	0.1875
Item 7 variance	0.2461	0.0588	0.0588
Item 8 variance	0.2644	0.0588	0.1094
Σ item variances	1.9262	0.4704	1.3260
Mean item variance	0.2408	0.0588	.1658
Variance of the mean item variances	0.0001	0.0000	0.0037
Total test variance	2.1094	0.5000	5.9963
Σ item variances/total test variance	0.9132	0.9408	0.2211
Alpha coefficient	0.0974	0.0714	0.8830
(Standard error of measurement)^2	2.0309	0.4952	0.7483
Relative mean item variance index	-0.4524	0.6454	
Relative squared standard error of estimate index	-1.7141	0.3382	

Note. All variance components were computed with the population size (N) and not the corrected sample size (N-1).